Divide-and-conquer based Summarization Framework for Extracting Affective Video Content

Irfan Mehmood[1], Muhammad Sajjad[1], Seungmin Rho[2], Sung Wook Baik[1*]

[1]College of Electronics and Information Engineering, Sejong University, Seoul, Republic of Korea

[2]Department of Multimedia, Sungkyul University, Anyang, Republic of Korea

irfanmehmood@sju.ac.kr, sajjad@sju.ac.kr, smrho@sungkyul.ac.kr, sbaik@sejong.ac.kr

Abstract: Recent advances in multimedia technology have led to tremendous increases in the available volume of video data, thereby creating a major requirement for efficient systems to manage such huge data volumes. Video summarization is one of the key techniques for accessing and managing large video libraries. Video summarization can be used to extract the affective contents of a video sequence to generate a concise representation of its content. Human attention models are an efficient means of affective content extraction. Existing visual attention driven summarization frameworks have high computational cost and memory requirements, as well as a lack of efficiency in accurately perceiving human attention. To cope with these issues, we propose a divide-and-conquer based framework for an efficient summarization of big video data. We divide the original video data into shots, where an attention model is computed from each shot in parallel. Viewer attention is based on multiple sensory perceptions, i.e., aural and visual, as well as the viewer's neuronal signals. The aural attention model is based on the Teager energy, instant amplitude, and instant frequency, whereas the visual attention model employs multi-scale contrast and motion intensity. Moreover, the neuronal attention is computed using the beta-band frequencies of neuronal signals. Next, an aggregated attention curve is generated using an intra- and inter-modality fusion mechanism. Finally, the affective content in each video shot is extracted. The fusion of multimedia and neuronal signals provides a bridge that links the digital representation of multimedia with the viewer's perceptions. Our experimental results indicate that the proposed shot-detection based divide-and-conquer strategy mitigates the time and computational complexity. Moreover, the proposed attention model provides an accurate reflection of the user preferences and facilitates the extraction of highly affective and personalized summaries.

Keywords: affective content analysis, big video data, divide-and-conquer-architecture, human attention modeling.

1. Introduction

Traditionally, computing has been mainly restricted to manipulating numerical and textual data. However, due to recent advances in digital systems, digitalized representations of images, audio, and video data have been introduced [1]. These heterogeneous media types and their combination in multimedia are a source of inspiration for the

*Corresponding Author

development of vast numbers of applications. At present, multimedia computing has been successfully incorporated into various applications such as video on demand, video conferencing, multimedia surveillance, and context-aware advertising. However, heterogeneous streams of data, high storage volumes, processing costs, and communication requirements demand a system that can manipulate such data in an efficient and effective manner. Experimental observations have shown that the data analysis and management algorithms used in traditional systems are not sufficient in assisting users when indexing and accessing their required content. To address this problem, video summarization schemes have been proposed that generate a concise version of a full-length video sequence by identifying the most important and pertinent content [2-5].

Video abstracts can be created in various forms, but these forms can generally be represented as keyframes or video skims [6]. Sets of keyframes, which are also called static storyboards, represent the main affective content of a video sequence using a group of salient frames. Video skimming is the extraction of a video clip with a much shorter duration than the original video sequence. The ultimate aim of video summarization is to algorithmically engineer computers to conceive specific multimedia data by giving the computers the ability to interpret multimedia data in the same manner as human perception. This helps content-based retrieval systems efficiently index content before being stored in a database. As a consequence, users can access their desired content more efficiently and effectively. Video abstracts can be generated manually and automatically, but a manual summarization is not feasible owing to the huge volumes of video data and limited manpower. Thus, the development of fully automated video analysis and processing algorithms is vital for reducing the manual involvement in the video summarization process [7]. Previous research suggests that existing video summarization schemes can be categorized into two classes: low- and high-level summarization schemes [8].

The majority of video summarization research during the last two decades has focused on the development of low-level video summarization techniques, which summarize video sequences by analyzing low-level features such as the color, shape, object motion, and speech [9]. For example, Naveed et al. [10], Zhou et al. [11], Avila et al. [12], Furini et al. [13], and Almeida et al. [14] use low-level features to summarize video rushes. Naveed et al. [10] use an aggregation mechanism to combine visual features, where keyframes were extracted from a video sequence based on the correlations among the RGB color channels, color histogram, and moments of inertia. Zhou et al. [11] initially extract the audio, color, and motion features, which are then dynamically fused using an adaptively weighting mechanism. The video sequence is clustered into separate scenes using a fuzzy c-means scheme with an optimally determined cluster number. Avila et al. [12] presented a scheme based on color feature extraction from video frames and a clustering algorithm for producing static video summaries. Initially, the frames are grouped in a sequential order instead of being randomly distributed between clusters. The frames are then grouped using the traditional k-means algorithm, leading to the generation of a summary selecting one frame per cluster. In [13], a summarization technique was specifically proposed for producing on-the-fly video storyboards. This method produces still and moving storyboards, which allows further advanced customization by users. This method is based on a fast clustering algorithm, which selects the most descriptive visual frames based on the HSV frame color distribution. For each frame in the input sequence, the visual features are extracted to describe the visual content.

After extracting the features, a simple and fast algorithm is used to detect groups of video frames with similar content and to select a representative frame from each group. Almeida et al. [14] segmented the input video into a set of meaningful shots by analyzing the color histogram of the image stream. Next, in each video shot, a Zero-mean Normalized Cross Correlation metric [15] is employed to eliminate redundant frames. Finally, the selected frames are filtered to avoid the inclusion of any possibly meaningless frames in the video summary.

Summarization schemes based on low-level features fail to agree with human perception owing to the semantic gap between low-level features and the high-level perception capability of humans with respect to video content [16]. To address this issue, most recent summarization schemes employ the concept of a visual attention model to bridge this semantic gap [17-19]. The basic assumption of such techniques is to extract frames as keyframes when they are visually important for humans as determined by visual attention models. Thus, the semantic details of video sequences can be better approximated compared with the low-level features. Attention is a supportive mental process in cognition and allows humans to interact with the outer world in a more focused and specialized manner [1]. Ma et al. [20] proposed the first attention-model based video summarization framework, which decomposes an original video sequence into the primary elements of its basic channels. Next, a set of features related to visual, aural, and linguistic attention is extracted to generate a comprehensive attention curve, which is used as an importance ranking, or to index the video content. Peng and Xiao-Lin [21] proposed a keyframe-based video summary method that uses visual attention cues, where static and dynamic attention models are computed and then fused using a motion priority scheme. This method controls the keyframe density according to the content variation in the entire clip. In [17-19], we previously proposed various visual-attention based summarization and prioritization schemes. These schemes employ the concepts of multiscale image contrast, salient motion, and linear and nonlinear weighted fusion methods to construct efficient human-attention models. Researchers have shown that visual attention-driven summarization techniques tend to obtain semantically more significant summaries compared with traditional low-level feature-based schemes. However, these attention models are computed using active features such as audio and visual information, whereas they ignore the passive responses of users, such as their neuronal signals. In addition, the complete human-attention mechanism is unknown. As a result, human-attention modeling algorithms fail to agree with the actual human-cognitive process.

The affective and emotional preferences of a user can overcome the drawbacks of existing visual and audio attention models, thereby generating an enhanced human-attention model [22]. An affect is generated through a neuronal process, which is triggered by the conscious/unconscious perception of a scene. An ideal summarization framework should consider both cognitive and emotional preferences. Emotional preferences can be measured by understanding the intensity and type of affects evoked in a user while watching a video sequence [23]. Various video scenes stimulate neuronal responses in the viewer's brain and such responses can be measured through an electroencephalograph (EEG) [24]. An EEG can provide valuable insight into the real-time changes in a viewer's affective state [25]. Owing to the technological advances, biosensors are now becoming more affordable, but they

are also more versatile in terms of capturing physiological response data [9]. For example, an EMOTIV EPOC headset[1] can measure a viewer's neuronal responses in real time.

To accurately understand a viewer's attentive and emotional preferences, we present an efficient human-attention model that combines both external (multimedia content) and internal information (the viewer's neuronal responses). Multimedia content-based features are extracted using audio and video processing, and neuronal attention features are extracted from EEG recordings. The multimedia and neuronal features are then combined to obtain an aggregated attention curve. This attention curve represents the inferred changes in the viewer's affective state while watching video sequences by extracting the most significant video frames. However, attention-driven frameworks to summarize large-scale video data are infeasible owing to their time and computational complexity. Moreover, an analysis of lengthy video sequences is impractical on a single computer because their data sizes are too large to store in memory. Thus, to increase the applicability of attention models in practical scenarios, we adopt a divide-and-conquer strategy that partitions the original video sequence into smaller sub-video clips. Our key idea for the dividing step is based on a shot-detection method, which detects shot boundaries using the histogram difference metric, $\mathcal{X}^2$. Next, attention models are computed and summaries are extracted from each sub-video clip independently, thereby achieving a much faster computational procedure. For the conquering step, we aggregate all sub-summaries and remove redundancies among them to build a final summary for the original video. This study offers three main contributions: 1) a shot-boundary detection based divide-and-conquer approach is presented to extract keyframes from each video shot independently, which is more promising than a sequential approach; 2) the viewer's neuronal responses are employed as a potential source of information; 3) and an efficient intra- and inter-modality attention fusion method is proposed for video summarization.

2. Methodology

Advances in multimedia technologies have produced a dramatic increase in the amount of video data. These data volumes exceed the capabilities of conventional video analytical methods. This situation demands algorithms capable of conducting an efficient analysis of large video data sets. In this context, we present a divide-and-conquer based framework to compute video summaries. The proposed framework splits input videos into a number of smaller video clips (video shots), extracting attention features from each of them independently, and merging the sub-summaries into a final summary, as shown in Figure 1. In each video shot, sub-summaries are extracted by computing the attention curves. Attention is considered to be the most important cognitive process associated with the human brain (e.g., reasoning, decision making, and problem solving). Thus, an efficient human attention model can play a vital role in an affective content analysis. The proposed attention model is computed using three types of data streams: audio, visual, and neuronal signals. Figure 2 shows the main steps in the proposed attention model.

[1] https://emotiv.com/epoc.php

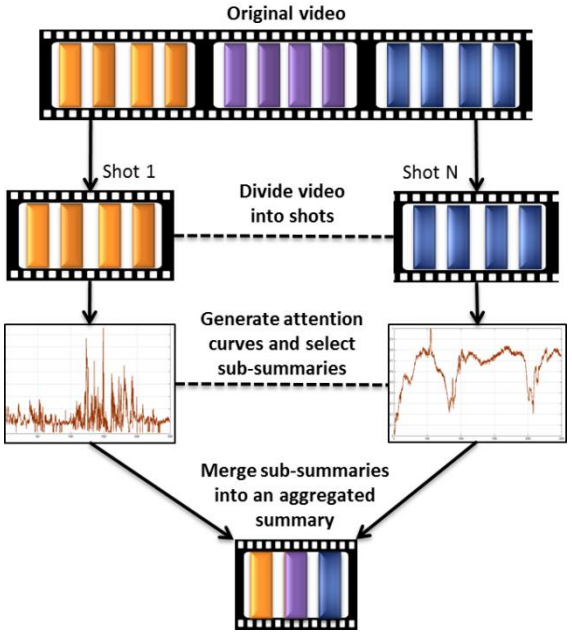

Figure 1. Conceptual diagram of the proposed divide-and-conquer based summarization process.

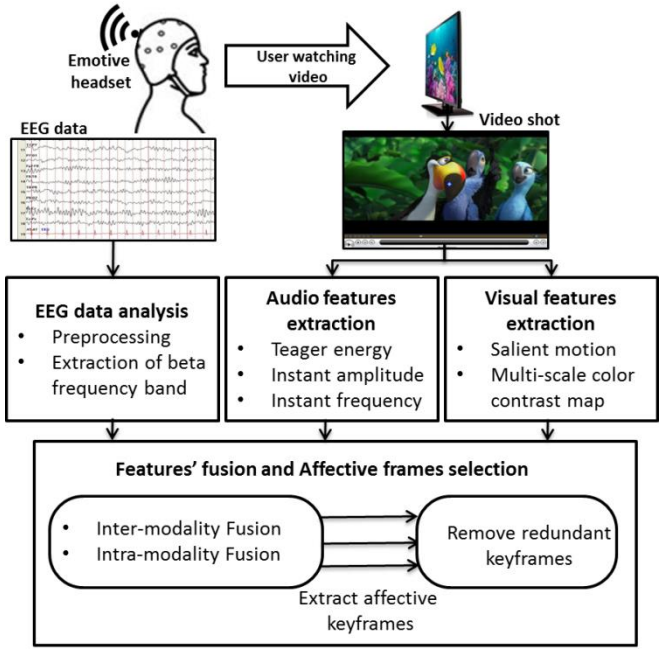

Figure 2. Framework of the proposed attention model for affective video content extraction.

2.1. Video Parsing

The analysis of lengthy video sequences is often infeasible owing to resource constraints and time complexity. In this context, video parsing can play a major role by dividing a long video into number of shorter video chunks. Amongst the various structural elements, a shot is an important temporal component of a video sequence. A video shot is a sequence of frames captured from a single camera operation. Shot detection, also known as video parsing,

is the first step in a video analysis. In video parsing, scene boundaries and scene changes between camera shots are detected [26], thereby making large amount of video data more manageable by imposing a hierarchy. A shot boundary is detected by identifying the transition (boundaries) between two consecutive shots. Shot boundaries are categorized into two classes of transitions: 1) abrupt (discontinuous) transitions, which are also referred to as cuts, and 2) gradual (continuous) transitions such as fades and dissolves, as shown in Figure 3. In recent years, various video parsing methods have been developed [27-29], and in a majority of shot-detection methods, the difference between successive frames is identified by comparing some of the low-level features such as the color histogram, edge, and intensity. These techniques are computationally efficient and accurately detect abrupt scene changes. However, they fail to detect gradual transitions within a scene. To cope with this challenge, we present a shot detection approach that partitions the video frames into 8×8 equally sized blocks and compares the corresponding blocks from two consecutive frames. This comparison is based on the $\mathcal{X}^2$ histogram measure, which is very effective in detecting both abrupt and gradual scene changes [30]. The $\mathcal{X}^2$ histogram difference between two frames $\mathcal{F}$(t) and $\mathcal{F}$(t+1) is computed as

$$d(\mathcal{F}(t),\mathcal{F}(t+1))_s = \sum_{k=1}^{n} \frac{\left(H_t^B(k) - H_{t+1}^B(k)\right)^2}{H_t^B(k)} \qquad (1)$$

where n is the number of histogram bins, and H_t^B and H_{t+1}^B are the color histograms of the Bth blocks of video frames $\mathcal{F}$(t) and $\mathcal{F}$(t+1), respectively. Video shots are easy to analyze with limited resources. In addition, it can reduce the time complexity and computational burden by allowing parallel and distributed processing of the underlying shots. Thus, a large summarization task can be divided into a number of subtasks, which is easy to achieve. After computing the local summaries from these shots, the sub-summaries can be efficiently merged to generate a final summary.

Figure 3. Examples of gradual and abrupt scene changes

2.2. EEG-based Attention Modeling

An EEG is used to detect electrical activity of the brain using electrodes attached to the user's scalp. An EEG generates a continuous recording of waves of varying frequencies and amplitudes. The number of EEG cycles per second is known as the frequency, which is recorded in Hertz (Hz). The amplitude is the strength of the EEG signal, which is measured in terms of microvolts of electrical energy. In general, EEG signals can be classified into five

categories according to their frequency bands, i.e., alpha, beta, theta, delta, and gamma bands. Researchers have found that activity in different EEG frequency bands can be related to specific psychophysiological states [31]. For instance, beta-band activity plays an important role in cognitive functions, mainly because of their relationship to the attention processes [32]. Thus, the beta-band (12 to 30 Hz) captures activities related to awakeness, alertness, and attentive states of mind. An increase in beta waves reflects the arousal of human attention. Thus, the beta band is the best option among the available EEG bands for generating a human attention model.

a. EEG Data Preprocessing

A wireless EMOTIV EPOC headset[2] can be used to capture EEG signals, which comprises 14 channels: AF3, AF4, F3, F4, F7, F8, FC5, FC6, P7, P8, T7, T8, O1, and O2. The captured EEG signals are usually contaminated with various artifacts caused by electrode movement, electrical line noise, muscle activity, sweating, and so on. Therefore, the EEG signals are preprocessed prior to an EEG analysis and feature extraction. Preprocessing of the EEG signals involves two steps: bandpass filtering and an independent component analysis (ICA). The biological artifacts caused by a heartbeat, blinking, and eyeball movements appear at around 1.2 Hz and below 4 Hz. Artifacts caused by muscle movements are the most dominant around 50 Hz. The Butterworth filter with a bandpass between 5 Hz and 50 Hz is applied to raw EEG signals to remove biological artifacts. This Butterworth filter rejects the unwanted frequencies, but also has a uniform sensitivity to the desired frequencies. In general, EEG recordings are linear combinations of the underlying brain sources [33]. As a result, the underlying signals become mixed during the EEG recordings. It is important to separate and extract each source by employing blind source separation (BSS) techniques [34]. Among the BSS techniques, ICA is known to be capable of estimating mutually independent sources from highly correlated EEG sources [35]. The main goal of ICA is to separate the unwanted artifacts from the neuronally generated EEG sources. In this study, we utilized EEGLAB[3] for processing and evaluating the ICA results based on the captured EEG data.

b. Extraction of EEG Features

Neurobiologists have obtained abundant evidence that attention is governed by the arousal level [36]. Arousal is the state of physiological reactivity in humans, which includes excitement, panic, and anger. In recent studies, psychologists have found that beta-band EEG frequencies are related to the arousal modulation in human cognitive processes. In various experiments, it has been shown that elderly people exhibit decreased beta-band power levels during tasks requiring visual attention [31]. This affect is accompanied by low behavioral accuracy in elderly people. It was thus determined that an increase in beta power indicates high attention, and vice versa. These findings are employed in the present study, and the power spectral densities (PSDs) of the EEG beta-band in each of the 14 channels are extracted as attention features. PSD is a positive and real function of the frequency variable associated with a stationary stochastic process, which describes the signal power strength at each frequency [37]. The PSD can

[2] https://emotiv.com/epoc.php
[3] http://sccn.ucsd.edu/eeglab/

be computed easily using a discrete Fourier transform (DFT). If we consider a 1-s non-overlapping EEG segment, its DFT coefficient $\mathcal{F}^c(f,t)$ at frequency *f* and time *t* is

$$\mathcal{F}^c(f,t) = \sum_{n=0}^{N-1} x(n,t) * e^{\left(-j\frac{2\pi}{N}kn\right)}; k = 1,2,3..., N-1 \quad (2)$$

where *x(n,t)* represents the discrete samples of EEG data at time *t* extracted from channel *c*, and *N* = 128 is the length of the EEG data per second. The PSD is calculated by taking the square of the absolute value of $\mathcal{F}^c(f,t)$ within the range of the beta-band frequencies, as follows:

$$\mathcal{PSD}_\beta^c(t) = \sum_{f=12}^{30} \left| F^c(f,t) \right|^2 \quad (3)$$

where f∈[12 30] because we consider only those frequencies that come under the category of a beta-band. The 14 features obtained through an EEG correspond to the data extracted from 14 microelectrodes: AF3, AF4, F3, F4, F7, F8, FC5, FC6, P7, P8, T7, T8, O1, and O2. The values of these features were normalized within the range of 0 to 1. We divided the PSD feature vectors into two classes: high- and low-priority channels, as shown in Figure 4. Four channels, i.e., P7, P8, O1, and O2, are included in the high-priority category, whereas the remaining 10 channels are considered to be of low-priority. The classification of features into high- and low-priority categories is motivated by the fact that visual attention is associated more deeply with responses in the occipital regions of the brain compared with non-occipital regions [38, 39]. Thus, the four channels that belong to the occipital regions, i.e., P7, P8, O1, and O2, are more important for the modeling of human attention. In this context, a weighted linear mechanism was used to estimate the EEG attention curve, as follows:

$$\mathcal{A}_{\mathrm{E}}(t) = \sum_{c \in LP} \mathcal{PSD}_\beta^c(t) * \mathcal{W}_{LP} + \sum_{c \in HP} \mathcal{PSD}_\beta^c(t) * \mathcal{W}_{HP}\ ;\ \mathcal{W}_{HP} > \mathcal{W}_{LP} \quad (4)$$

where W_{LP} and W_{HP} are the weights of the low- and high-priority channels, respectively. The weight values are within the range of 0 to 1. To give more weighting to high-priority channels, the value of W_{HP} was kept higher than that of W_{LP}. In addition, $\mathcal{A}_E(t)$ is an EEG-based attention curve, which represents the attention level of the viewer at time *t* while watching a video sequence. This attention curve was normalized in the range of 0 to 1. We carefully synchronized the visual output of the video sequences used with the viewers' EEG data recordings; thus, the EEG feature values denote the attention levels for the corresponding video frames. When the attention value was close to 1, the frame sequence viewed at a particular time was considered to be salient (i.e., affective) for that user. Similarly, when the attention value was close to 0, the viewed frame sequence was considered to be non-salient (i.e., non-affective).

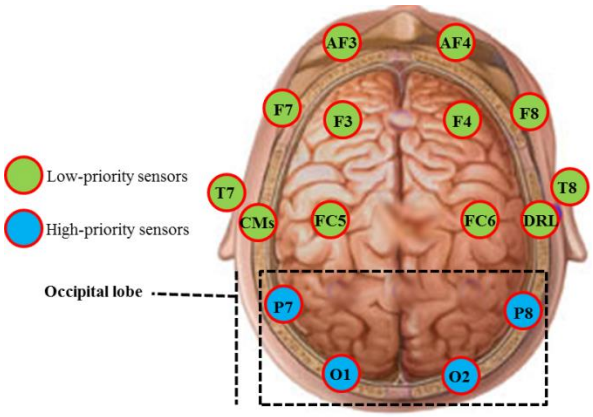

Figure 4. Top view of a human brain model showing the EEG electrode locations, as Traced from [40].

2.3. Audio Visual Attention Modeling

Researchers have identified significant correlations between audiovisual information and human emotional states [41]. Some of the most important components of visual information are salient motion (rhythm motion) and multiscale image contrast. Rhythm motion measures the amount of motion in video frames. The quantity of the salient motion present in a particular part of a video frame is directly proportional to its level of excitement. Similarly, multiscale image contrast is an essential visual feature for attention detection because the contrast operator simulates the human visual receptive field. Multi-scale contrast is used to compute the local contrast features of intensity at each scale. We used these features to produce compelling video summaries, as described in our previous schemes [17, 42, 43], and they are integrated into our current framework. Audio signals also provide useful information when selecting more affective and semantically meaningful video content [44, 45]; therefore, in addition to a visual content analysis, audio features are also incorporated in the proposed framework. In [46], an AM-FM modulation model was proposed for speech, which allows speech formants to be modeled as follows:

$$S(t) = \sum_{N=0}^{N-1} a_N(t) * \cos(\varphi_N(t)) \tag{5}$$

where $a_N(t)$ is the amplitude, $\Phi_N(t)$ is the frequency, and N is the size of the audio sequence. The audio attention model is based on three features: the maximum Teager energy, M_{TE}: the mean instant amplitude, M_{IA}: and the mean instant frequency, M_{IF} [47]. The first, M_{TE} captures the joint amplitude-frequency information of the audio activity, which represents the dominant signal modulation energy. For an audio signal frame m of length N, M_{TE} is obtained by measuring the Gabor filter bank responses on S as:

$$\mathcal{M}_{\text{TE}}(m) = \max_{1 \le k \le K} \frac{1}{N} \sum_{n}^{N} \Psi_d\left((S * h_k)[n]\right); \ (m-1)N \le n \le mN \tag{6}$$

where $*$ is the convolution operator, ψ_d is the nonlinear Teager-Kaiser differential energy operator, K represents the linearly spaced bank of Gabor filters, h_k is the impulse response of the k^{th} filter, and n is the sample index. Next, an energy separation algorithm [48] is applied on filter j = arg max (M_{TE}) to derive the M_{IA} and M_{IF} features:

$$\mathcal{M}_{\mathrm{IA}}(m) = \left(\overline{A_j[n]}\right) \tag{7}$$

$$\mathcal{M}_{\mathrm{IF}}(m) = \left(\overline{\Omega_j[n]}\right) \tag{8}$$

where A and Ω are the instantaneous amplitude and frequency signals, respectively. An aural attention curve is then generated by linearly fusing the M_{TE}, M_{IA}, and M_{IF} audio features to signify the salient audio segments.

2.4. Extracting Keyframes from Intra- and Inter-Modality Attention Models

The extraction of high-level neuronal features and low- or middle-level multimedia features allows the extraction of keyframes. Figure 5 shows the intra- and inter-modality fusion and keyframe selection process. At the intra-modality level, keyframes are extracted independently using attention features based on the modality at the time using the aural, visual, and EEG attention curves computed as described in the previous section. The keyframes selected by each modality are then combined to generate a final intra-modality summary. However, there is a semantic gap between the high-level and low-level features due to the lack of an appropriate description and the presentation of the semantics perceived by the human brain. To overcome these issues, the inter-modality attention model is computed by linearly fusing audio, visual, and EEG-based attention curves. This inter-modality attention fusion method combines the strengths of the underlying three modalities while minimizing their weaknesses. The summaries generated by intra- and intra-modality are merged to generate a final summary, but it was observed that the aggregated summary has some redundant frames. Therefore, these redundant frames are removed using a color histogram.

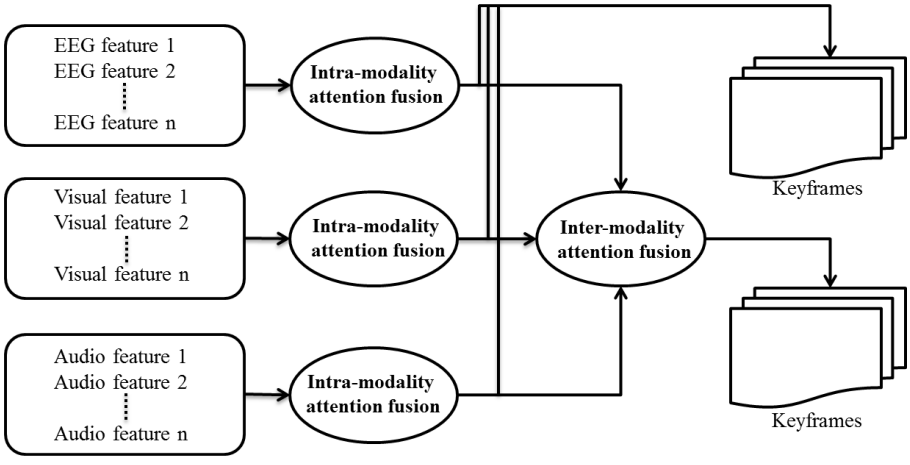

Figure 5. Keyframe extraction from intra- and inter-modality attention models.

3. Experiments and Results

To evaluate the efficacy of the proposed method, various experiments were conducted using video sequences from different genres. The details of these experiments are described in the following sections.

3.1. Experimental Setup

In this subsection, we describe the technical setup, participants, ambient environment, and goals of the experiments. The main goal of this experiment was to assess the ability of the proposed algorithm to correctly detect the salient frames in a video sequence. Tests were conducted using numerous videos downloaded from two standard databases: *The Open Video Project*[4] and *AiirSource channel*[5]. The selected video sequences belonged to different genres, including action, thriller, horror, and comedy. Detailed information regarding the sample video sequences used in this experiment is shown in Table 1. Ten users (five female and five male) from different laboratories in the Digital Content Department, which is close to our laboratory, were requested to participate in this study. All of the participants have normal vision and reported no history of neurological problems. In addition, they are all researchers working on image analysis. The experiments were conducted in a fully equipped multimedia laboratory, which was free from outside noise. The participants were provided two 22-inch displays, which facilitated their high resolution viewing of the videos. The visual field of each participant was protected from outside distractions, thereby providing a clear view of the screen. The output volume of the speakers was also tested to ensure that the users could clearly hear the audio content of the video clips.

The EEG signals of each participant were recorded using an EMOTIV EPOC headset[6] while watching the videos. The EMOTIV detected and digitized the EEG signals produced by each participant's brain and wirelessly transmitted them to a computer. We used the TestBench™ research software included in the EMOTIV Research Edition SDK, which facilitates the recording of EEG data files in the binary EEGLAB format. The main use of the EEG recordings was for extracting the neuronal responses of each participant and to investigate these responses to select the most salient contents of the underlying video. Before initiating the experiment, the participants were asked to wear the EMOTIV headset for a few minutes to familiarize themselves with the sensors. This avoided potential discomfort during use, which could have influenced the accuracy of the experimental results. After watching each video clip, the participants were asked to provide a summary of the underlying video, where they selected the most informative and affective frames. The summaries produced manually by the participants were used as the ground truth of the comparisons. The EEG recordings were carefully synchronized with the audio and visual frames. EEGLAB's toolbox was used to process the acquired EEG data, whereas Matlab was used to extract the audio and visual features.

Table 1: Details of the test video database.

No.	Video title	Video description	Genre
1	Skydiving	A group of youngsters videotape their skydiving trip. One person's parachute does not open and he has to use his reserve chute.	Adventure, Fear

[4] http://www.open-video.org/index.php
[5] https://www.youtube.com/user/AiirSource
[6] https://emotiv.com/epoc.php

2	Rave: Christmas Special	This video shows a joyful holiday season for many, except Vernon who is constantly bullied. However, the Christmas trees come alive and take revenge on his behalf.	Fantasy, Comedy
3	Bunjee	People bungee jumping from a bridge in New Zealand.	Adventure, Thrill
4	Winning: Aerospace, Segment 07	This video introduces students to the unique career opportunities in America's aerospace industry.	Education, Documentary
5	1955 Chevrolet Screen Ads	Ten short theatrical "screen ads" promoting 1955 Chevrolet models.	Adventure, Lifestyle
6	NASASciFiles - Aviation History	NASA science file segment tracing the history of human flight.	Documentary, Thriller
7	U.S. Marines Maritime Raid Force - MH-60S Helicopter Casting and SPIE	U.S. Marines Maritime Raid Force conducting helicopter casting and helicopter special patrol operations from an MH-60S Helicopter	Adventure, Thriller
8	U.S. Marines Pilot Recovery Training. HH-60 Pave Hawk	A squadron and a group of U.S. Marines visiting Eielson Air Force Base, Alaska to execute pilot recovery training	Adventure, Thriller
9	US Airstrike Against ISIS Storage Facility in Syria	Unmanned aerial vehicle video of overnight operations against ISIL in Syria, including a U.S. airstrike against an ISIS/ISIL storage facility near Abu Kamel, Syria.	Action, War
10	Drone and Manned F-18 Takeoff and Land	The US Navy's unmanned X-47B conducts flight operations with a manned F-18 aboard the aircraft carrier, the USS Theodore Roosevelt (CVN 71).	Action, Thriller

3.2. Case Study: Extracting Keyframes from a Single Video

In this subsection, we demonstrate the benefits of the proposed scheme based on the choice of keyframes in the video sequence *Daenerys' Dragons Fight*[7] taken from the American medieval fantasy television series *Game of Thrones*. This video sequence comprises of 2490 frames, which capture Daenerys Targaryen watching a fight between her dragons. Daenerys Targaryen, who has the sobriquet Mother of Dragons, is one of the major characters in this series. In the video sequence, one of the dragons unexpectedly snaps at Daenerys when she tries to interfere with the dragons that are fighting over food; thus, she realizes that she is losing control of them.

For the underlying video, the intra-modality audio, visual, and EEG-based attention curves are shown in Figures 6(a), 6(b), and 6(c), respectively. Figure 6(a) shows the audio attention curve estimated by fusing three audio attention features, i.e., the maximum Teager energy, the mean instant amplitude, and the mean instant frequency. Figure 6(a) shows that the values of the audio attention curve did not vary significantly during the initial 30s of the video sequence. However, a significant increase in the audio attention was observed during the first 30 to 60s, which was caused by the horrifying sounds of the fighting dragons. In the last 23s, a decrease in audio attention was observed with only a few significant sounds. The keyframes extracted using the audio attention curve are shown in Figure 7(a). Figure 6(b) shows the visual attention curve. In frames 730 through 828, no salient activities were observed. Consequently, visual attention received the minimum value for this sequence. In contrast, a significant increase in visual attention was observed in frames 300 through 400, which shows Daenerys and her pair of dragons fighting in the sky. Figure 7(b) shows the keyframes extracted from the underlying video using the visual attention curve. Figure 6(c) shows the EEG power spectrum of a viewer while watching the video sequence. In this study, the EEG attention curve was found to be more suitable for locating exciting video segments, as shown in Figure 7(c).

[7] https://www.youtube.com/watch?v=RHTw8BxU1fk

Frames 573 and 1655 showed positive and negative events that occurred in the video, respectively. These frames were efficiently highlighted by the beta band-driven EEG attention curve.

A few important frames were neglected by the intra-modality attention models. To overcome this issue, an inter-modality attention model was estimated by linearly fusing various modalities, i.e., the audio, visual, and EEG attention curves, to generate a single attention curve, as shown in Figure 6(d). This inter-modality attention curve overcame the weaknesses of the three underlying modalities. Figure 7(d) shows the summary generated from the inter-modality fusion curve. The summaries generated by the intra- and inter-modality models were combined to generate a final video summary. Redundant frames were eliminated from the aggregated summary using a color histogram. The video summary obtained matched the human perception by representing more interesting and representative frames from the video sequence, as shown in Figure 7(e).

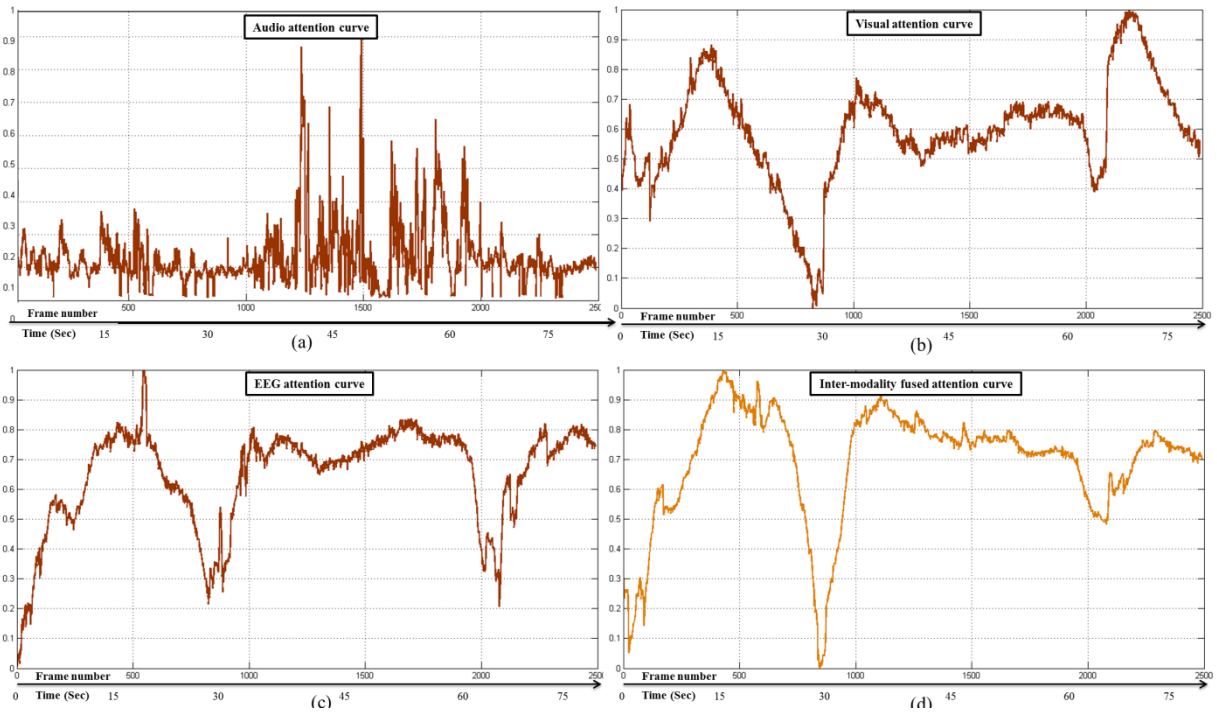

Figure 6. Intra- and inter-modality attention curves for the video sequence *Daenerys' Dragons Fight*.

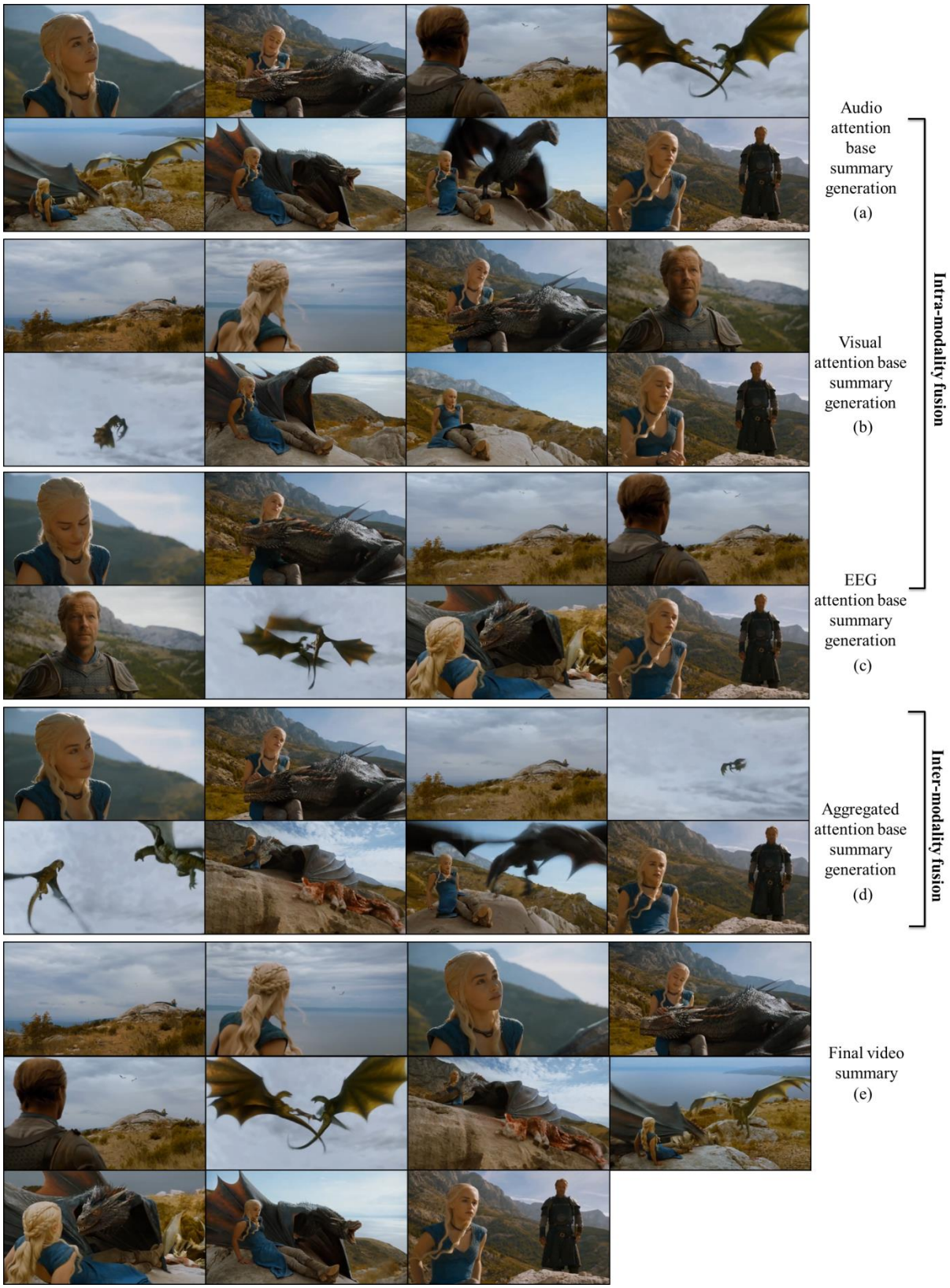

Figure 7. Comparison of the affective video contents extracted by intra- and inter-modality attention models.

3.3. Ground Truth Comparison based Summary Evaluation

Ground truth comparison-based evaluation metrics are widely used to assess the effectiveness of video summarization schemes. In this study, the ground truth was prepared by a group of users and the summaries produced by a specific technique were compared with the ground truth to compute the values of various evaluation metrics. The ground truth data for the dataset in Table 1 were manually generated by a group of multimedia experts. In this section, we describe the efficiency of the proposed method based on two sets of metrics: 1) precision, recall, and, f-measure metrics, and 2) the TRECVID evaluation metrics.

a. Evaluations based on the Precision, Recall, and F-measure

The precision, recall, and f-measure are widely employed metrics for evaluating video summarization and video retrieval tasks [49-51]. In a video summarization, precision is the ratio of the number of relevant frames chosen as keyframes by the system to the total number of relevant and irrelevant frames selected as keyframes by the system. Recall is the ratio of the number of relevant frames chosen as keyframes by the system to the total number of keyframes in the ground truth summary. Precision and recall are complementary metrics and cannot be used solely. For example, a high precision value can be obtained by generating a short summary, which includes few relevant frames. Similarly, extracting too many keyframes can lead to a high recall value. The f-measure is the average of the precision and recall, which facilitates an interpretation of the results by providing a combined measure. A high F-measure indicates that both the precision and recall have high values. Precision, recall and f-measure are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$\mathcal{F}\text{-}measure = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{11}$$

where a true positive (TP) refers to the number of frames extracted by the summarization scheme that are also present in the ground truth. A false positive (FP) refers to the number of frames selected by the scheme that are not present in the ground truth. A false negative (FN) is defined as the number of frames included in the ground truth that are not selected by the summarization scheme. The precision, recall, and f-measure range from 0 to 1. A value close to 1 is considered good, and vice versa. Table 2 shows the mean precision, recall, and f-measure values obtained by the proposed method using different attention fusion schemes. The proposed aggregated attention model obtained significant improvements compared with the summaries generated by individual intra- and inter-modality attention models. In some cases, the audio, visual, and EEG attention models obtained high values for one of the metrics. However, a high value for one of the metrics is generally insufficient. For example, in video sequence 7, the

visual-attention base summary achieved a precision value of 1 by selecting all of the relevant frames but the recall value was low. However, the aggregated summarization scheme had the highest Recall value for video 7 and a sufficiently high precision value. In addition, the f-measure of the aggregated summary was 0.77, whereas that for visual attention-based summary was 0.71. Table 2 shows that the overall aggregated attention model achieved high F-measure values.

Table 2: Mean recall (R), precision (P), and f-measure (F) values obtained by the proposed method using different fusion schemes.

	Intra-modality attention									**Inter-modality attention**			**Aggregated**		
	Audio			**Visual**			**EEG**								
No.	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
1	0.48	0.61	0.54	0.72	0.55	0.62	0.75	0.74	0.74	0.47	0.49	0.48	0.88	0.77	0.82
2	0.50	0.70	0.58	0.64	0.68	0.66	0.81	0.67	0.73	0.71	0.77	0.74	0.79	0.70	0.74
3	0.74	0.49	0.59	0.51	0.71	0.59	0.74	0.77	0.75	0.77	0.80	0.78	0.92	0.94	0.93
4	0.82	0.52	0.64	0.44	0.54	0.48	0.66	0.61	0.64	0.70	0.67	0.68	0.94	0.85	0.89
5	0.66	0.49	0.56	0.49	0.66	0.56	0.76	0.80	0.78	0.74	0.75	0.74	0.91	0.81	0.86
6	0.60	0.67	0.63	0.53	0.73	0.61	0.79	0.70	0.74	0.78	0.67	0.72	0.88	0.86	0.87
7	0.47	0.58	0.52	0.55	1.00	0.71	0.81	0.68	0.74	0.70	0.68	0.69	0.88	0.69	0.77
8	0.62	0.70	0.66	0.55	0.68	0.61	0.68	0.88	0.77	0.71	0.77	0.74	0.96	0.54	0.69
9	0.55	0.65	0.60	0.65	0.71	0.68	0.65	0.70	0.67	0.69	0.80	0.74	0.95	0.90	0.92
10	0.51	0.73	0.60	0.72	0.71	0.72	0.69	0.60	0.64	0.75	0.66	0.70	0.93	0.95	0.94
Average	**0.60**	**0.61**	**0.59**	**0.58**	**0.70**	**0.63**	**0.73**	**0.72**	**0.72**	**0.70**	**0.71**	**0.70**	**0.90**	**0.80**	**0.84**

b. TRECVID Evaluation Criteria

In this subsection, we present comparisons of the proposed scheme against state-of-the-art summarization schemes based on non-visual and visual attention, i.e., STIMO [13], VSUMM [12], and the method proposed by Naveed et al. [17]. This comparison used the TRECVID evaluation metrics. TRECVID is a conference series that promotes research into information retrieval by providing a large test data collection and efficient evaluation procedures [52]. TRECVID has defined its evaluation criteria based on comparisons of summaries with the ground truth. In the evaluation performed in the present study, a system-generated summary was assessed by a human user and compared with the ground truth summary. During this comparison, the user was requested to score the underlying summary based on three criteria: (1) the amount of ground truth included, (2) the amount of redundancy present, and (3) the proportion of junk frames present in the summary. The value of the ground truth inclusion (IN) was normalized within a range of 0 to 1, where a value close to 1 indicates that most of the frames in the ground truth were covered by the system-generated summary. The scores for a lack of redundancy (RE) and lack of junk frames (JU) were normalized within a range of 1 to 5, where a score close to 5 indicates the best scenario, with minimum redundancy and minimum junk frames; whereas a score close to 1 denotes that the summary contains redundant and junk frames. Table 3 indicates that the proposed method obtained a high value for the IN metric. In addition, a high score for the RE metric demonstrates that the summary evaluators (users) strongly disagree that the proposed summary contains redundancy in comparison with the ground truth. Our method removed junk frames in an efficient manner, and obtained high scores for the JU metric. There were some exceptions, but the proposed method generally outperformed the other summarization schemes. For instance, for video sequence 3, the proposed method achieves

high scores of redundancy and junk metrics, which highlights the significance of our method to remove useless and redundant frames. However, for this video, the proposed method's summary has a low inclusion score. Similarly, STIMO has the highest value of redundancy for video sequence 9, whereas the inclusion value is significantly low.

Table 3: Inclusion (IN), redundancy (RE), and junk frame (JU) rating for different schemes.

	Non-visual-attention based schemes						**Visual-attention based scheme**			**Proposed method**		
	STIMO [13]			**VSUMM** [12]			**Naveed et al.** [17]					
No.	IN	RE	JU	IN	RE	JU	IN	RE	JU	IN	RE	JU
1	0.69	4	3	0.88	4	4	0.74	4	3	0.84	4	5
2	0.65	3	3	0.77	3	3	0.69	4	4	0.81	4	5
3	0.78	5	4	0.63	3	4	0.88	5	4	0.77	5	5
4	0.65	4	3	0.55	2	4	0.71	4	5	0.92	3	4
5	0.62	4	4	0.72	5	5	0.7	3	4	0.79	4	5
6	0.68	4	3	0.7	4	4	0.66	4	3	0.88	4	5
7	0.64	4	3	0.59	3	2	0.69	3	4	0.8	5	5
8	0.57	3	4	0.62	3	3	0.83	4	2	1	4	4
9	0.52	3	3	0.66	3	4	0.62	3	3	0.9	4	4
10	0.64	3	4	0.79	4	4	0.72	4	3	0.74	5	5
Average	**0.64**	**3.7**	**3.4**	**0.69**	**3.4**	**3.7**	**0.72**	**3.8**	**3.5**	**0.84**	**4.2**	**4.7**

3.4. Comparative Analysis of Computational Time

To validate the effectiveness of the proposed summarization framework in terms of the computational time, we extracted video summaries using both sequential and divide-and-conquer based approaches. In the sequential process, summaries were extracted from video sequences without fragmentation, i.e., a complete video sequence was processed on a single server. In the divide-and-conquer based approach, the underlying video sequence was first fragmented into shots and then sub-summaries were extracted from each shot simultaneously. Figure 8, depicts the comparison between the total computational time required for a summarization of the underlying sample video database. It can be observed that total summary generation time using the sequential process is significantly high compared with the corresponding divide-and-conquer based process. Although the total CPU consumption time, which is the sum of the CPU usage time in computing sub-summaries on all nodes, was approximately equal to or higher than sequential processing, the distribution of video shots among several nodes for a parallel execution reduced the total summarization time. This result is significant when dealing with large video repositories.

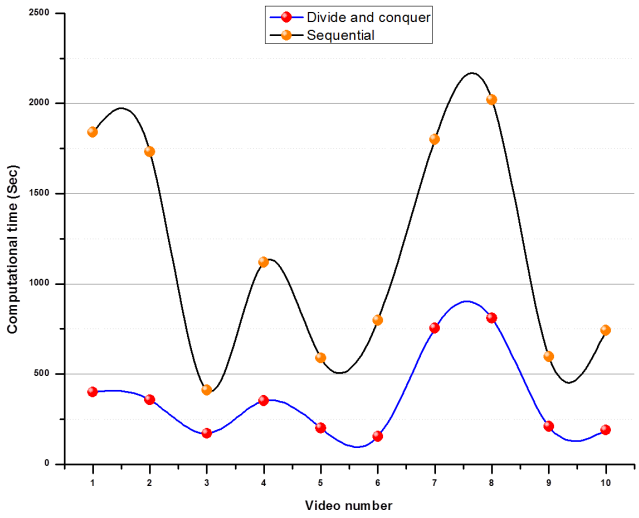

Figure 8. Comparison of summary generation time between a traditional approach and the proposed divide-and-conquer framework.

4. Conclusion

In this study, we proposed a divide-and-conquer based summarization framework, which computes aural, visual, and EEG attention features to extract the affective keyframes from video sequences. In the dividing step, audio and video streams are fragmented into shots, and each video shot is distributed among the nodes to compute the attention features. In the conquering step, sub-summaries generated from individual shots are combined to obtain the final summary. The viewer's attention is modeled based on multiple sensory perceptions, i.e., aural, visual, and neuronal signals. Aural attention is defined using three intrinsic attributes: the maximum Teager energy, the mean instant amplitude, and the mean instant frequency, whereas visual attention is computed from two essential video attributes, i.e., multi-scale contrast and salient motion. Aural and visual attention models are both related to active attention, and are often controlled by the video producers, e.g., the camera motions and sound level are controlled by the producer to reflect the producer's intentions. In addition to active attention, passive attention is computed using EEG signals. Our proposed EEG-based attention technique is based on the beta-band frequencies obtained from the neuronal signals of the viewers. We found that EEG attention helps to identify the preferences and emotions of users with respect to a particular video. Finally, all of the attention models obtained from different modalities are fused to generate an aggregated attention model. The keyframes selected by the aural, visual, and EEG attention models, as well as those based on the aggregated attention model, are combined to generate a final summary. We compared the performance of the proposed method with several state-of-the-art summarization schemes based on two standard sets of metrics: 1) precision, recall, and f-measure, and a 2) TRECVID evaluation, which demonstrates that the proposed framework selects semantically relevant keyframes from video sequences and personalizes the summary according to the viewer's preferences. We also presented experimental results that validate the effectiveness of the proposed divide-and-conquer framework, reducing the computational time and optimizing the resources by distributing the summarization tasks among various computation nodes.

Acknowledgment

This research is supported by (1) the ICT R&D program of MSIP/IITP. [2014(R0112-14-1014), The Development of Open Platform for Service of Convergence Contents], and (2) The Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2012904).

References

[1] H. Ekanayake, CognitiveEmotional User Correction for Multimedia Interactions Using Visual Attention and Psychophysiological Signals 2009.
[2] N. Ejaz, S.W. Baik, Video summarization using a network of radial basis functions, Multimedia systems 18 (2012) 483-497.
[3] A.G. Money, H. Agius, Video summarisation: A conceptual framework and survey of the state of the art, Journal of Visual Communication and Image Representation, 19 (2008) 121–143.
[4] J. Niu, D. Huo, K. Wang, C. Tong, Real-time generation of personalized home video summaries on mobile devices, Neurocomputing 120 (2013) 404–414.
[5] I. Mehmood, M. Sajjad, S.W. Baik, Mobile-Cloud Assisted Video Summarization Framework for Efficient Management of Remote Sensing Data Generated by Wireless Capsule Sensors, Sensors, 14 (2014) 17112–17145.
[6] B.T. Truong, S. Venkatesh, Video abstraction: A systematic review and classification, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), 3 (2007) 3.
[7] T. Wan, Z. Qin, A new technique for summarizing video sequences through histogram evolution, Signal Processing and Communications (SPCOM), 2010 International Conference on, (IEEE2010), pp. 1–5.
[8] A.C. Bovik, Handbook of Image and Video Processing, Academic Press, 2010.
[9] A.G. Money, H. Agius, Elvis: entertainment-led video summaries, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 6 (2010) 17.
[10] N. Ejaz, T.B. Tariq, S.W. Baik, Adaptive key frame extraction for video summarization using an aggregation mechanism, Journal of Visual Communication and Image Representation, 23 (2012) 1031–1040.
[11] H. Zhou, A.H. Sadka, M.R. Swash, J. Azizi, U.A. Sadiq, Feature extraction and clustering for dynamic video summarisation, Neurocomputing, 73 (2010) 1718–1729.
[12] S.E.F. de Avila, A.P.B. Lopes, A. da Luz, A. de Albuquerque Araújo, VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recognition Letters, 32 (2011) 56–68.
[13] M. Furini, F. Geraci, M. Montangero, M. Pellegrini, STIMO: STIll and MOving video storyboard for the web scenario, Multimedia Tools and Applications, 46 (2010) 47–69.
[14] J. Almeida, N.J. Leite, R.d.S. Torres, Vison: Video summarization for online applications, Pattern Recognition Letters, 33 (2012) 397–409.
[15] J. Crowley, J. Martin, Experimental comparison of correlation techniques, IAS-4, International Conference on Intelligent Autonomous Systems, Karlsruhe1995).
[16] W. Hu, N. Xie, L. Li, X. Zeng, S. Maybank, A survey on visual content-based video indexing and retrieval, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 41 (2011) 797–819.
[17] N. Ejaz, I. Mehmood, S.W. Baik, Efficient visual attention based framework for extracting key frames from videos, Signal Processing: Image Communication, 28 (2013) 34–44.
[18] I. Mehmood, M. Sajjad, S.W. Baik, Video summarization based tele-endoscopy: A service to efficiently manage visual data generated during wireless capsule endoscopy procedure, Journal of medical systems, 38 (2014) 1–9.

[19] N. Ejaz, I. Mehmood, S.W. Baik, Feature aggregation based visual attention model for video summarization, Computers & Electrical Engineering, 40 (2014) 993–1005.
[20] Y.-F. Ma, X.-S. Hua, L. Lu, H.-J. Zhang, A generic framework of user attention model and its application in video summarization, Multimedia, IEEE Transactions on, 7 (2005) 907–919.
[21] J. Peng, Q. Xiao-Lin, Keyframe-based video summary using visual attention clues, IEEE MultiMedia, (2009) 64-73.
[22] M.S. Irfan Mehmood, S. Rho, and Sung Wook Baik, Audio-visual and EEG-based Attention Modeling for Extraction of Affective Video Content, (PlatCon 2015, Jeju island, Korea., January 26–28, 2015).
[23] S. Zhao, H. Yao, X. Sun, Video classification and recommendation based on affective analysis of viewers, Neurocomputing, 119 (2013) 101–110.
[24] E. Niedermeyer, F.L. da Silva, Electroencephalography: basic principles, clinical applications, and related fields (Lippincott Williams & Wilkins, 2005).
[25] X.-W. Wang, D. Nie, B.-L. Lu, Emotional state classification from EEG data using machine learning approach, Neurocomputing, 129 (2014) 94–106.
[26] S. Tsekeridou, I. Pitas, Content-based video parsing and indexing based on audio-visual interaction, Circuits and Systems for Video Technology, IEEE Transactions on, 11 (2001) 522–535.
[27] J. Mas, G. Fernandez, Video shot boundary detection based on color histogram, Notebook Papers TRECVID2003, Gaithersburg, Maryland, NIST, (2003).
[28] R. Vázquez-Martín, A. Bandera, Spatio-temporal feature-based keyframe detection from video shots using spectral clustering, Pattern Recognition Letters 34 (2013) 770–779.
[29] M. Birinci, S. Kiranyaz, A perceptual scheme for fully automatic video shot boundary detection, Signal Processing: Image Communication (2013).
[30] H. Zhang, A. Kankanhalli, S.W. Smoliar, Automatic partitioning of full-motion video, Multimedia systems, 1 (1993) 10–28.
[31] M. Gola, M. Magnuski, I. Szumska, A. Wróbel, EEG beta band activity is related to attention and attentional deficits in the visual performance of elderly subjects, International Journal of Psychophysiology 89 (2013) 334–341.
[32] M. Gola, J. Kamiński, A. Brzezicka, A. Wróbel, Beta band oscillations as a correlate of alertness—Changes in aging, International Journal of Psychophysiology 85 (2012) 62–67.
[33] M. Hassan, O. Dufor, I. Merlet, C. Berrou, F. Wendling, EEG source connectivity analysis: From dense array recordings to brain networks, PloS one, 9 (2014) e105041.
[34] A. Jafarifarmand, M.A. Badamchizadeh, Artifacts removal in EEG signal using a new neural network enhanced adaptive filter, Neurocomputing 103 (2013) 222–231.
[35] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M.J. Mckeown, V. Iragui, T.J. Sejnowski, Removing electroencephalographic artifacts by blind source separation, Psychophysiology, 37 (2000) 163–178.
[36] J.T. Coull, Neural correlates of attention and arousal: Insights from electrophysiology, functional neuroimaging and psychopharmacology, Progress in Neurobiology 55 (1998) 343–361.
[37] S. Sanei, J.A. Chambers, EEG Signal Processing (John Wiley & Sons, 2008).
[38] A. Belyavin, N.A. Wright, Changes in electrical activity of the brain with vigilance, Electroencephalography and Clinical Neurophysiology, 66 (1987) 137–144.
[39] R. Ptak, F. Lazeyras, M. Di Pietro, A. Schnider, S.R. Simon, Visual object agnosia is associated with a breakdown of object-selective responses in the lateral occipital cortex, Neuropsychologia 60 (2014) 10–20.
[40] K. Sundararaman, J. Parthasarathi, S.V. Rao, G. Appa Rao, Hridaya A tele-medicine initiative for cardiovascular disease through convergence of grid, Web 2.0 and SaaS, Pervasive Computing Technologies for Healthcare, 2008. PervasiveHealth 2008. Second International Conference on, (IEEE2008), pp. 15–18.
[41] A. Hanjalic, L.-Q. Xu, Affective video content representation and modeling, Multimedia, IEEE Transactions on, 7 (2005) 143–154.

[42] I. Mehmood, M. Sajjad, W. Ejaz, S.W. Baik, Saliency-directed prioritization of visual data in wireless surveillance networks, Information Fusion (2014).
[43] I. Mehmood, N. Ejaz, M. Sajjad, S.W. Baik, Prioritization of brain MRI volumes using medical image perception model and tumor region segmentation, Computers in biology and medicine, 43 (2013) 1471–1483.
[44] S. Rho, E. Hwang, Video scene determination using audiovisual data analysis, Distributed Computing Systems Workshops, 2004. Proceedings on 24th International Conference (IEEE2004), pp. 124–129.
[45] D. Kim, D. Kim, S. Jun, S. Rho, E. Hwang, TrendsSummary: A platform for retrieving and summarizing trendy multimedia contents, Multimedia Tools and Applications, (2013) 1–16.
[46] A.C. Bovik, P. Maragos, T.F. Quatieri, AM-FM energy detection and separation in noise using multiband energy operators, Signal Processing, IEEE Transactions on 41 (1993) 3245-3265.
[47] G. Evangelopoulos, K. Rapantzikos, P. Maragos, Y. Avrithis, A. Potamianos, Audiovisual attention modeling and salient event detection, Multimodal Processing and Interaction, Springer, (2008) 1–21.
[48] P. Maragos, J.F. Kaiser, T.F. Quatieri, Energy separation in signal modulations with application to speech analysis, Signal Processing, IEEE Transactions on 41 (1993) 3024–3051.
[49] S. Rho, E. Hwang, FMF: Query adaptive melody retrieval system, Journal of Systems and Software 79 (2006) 43-56.
[50] C. Liu, Q. Huang, S. Jiang, L. Xing, Q. Ye, W. Gao, A framework for flexible summarization of racquet sports video using multiple modalities, Computer Vision and Image Understanding 113 (2009) 415–424.
[51] E. Dumont, B. Merialdo, Rushes video summarization and evaluation, Multimedia Tools and Applications 48 (2010) 51–68.
[52] P. Over, A.F. Smeaton, P. Kelly, The TRECVID 2007 BBC rushes summarization evaluation pilot, Proceedings of the International Workshop on TRECVID Video Summarization, ACM2007, 1–15.